LA SIGNIFICATION ÉPISTÉMOLOGIQUE DE LA THÉORIE DES AUTOMATES

Le développement de l'analogie entre information et entropie via l'expérience de pensée du démon de Maxwell représente un moment décisif dans la constitution cybernétique du concept d'information. La cybernétique a conçu l'étude du traitement de l'information comme une branche de la mécanique statistique. Cette relation originale d'inclusion de concepts issus de la théorie des télécommunications au sein de la physique est porteuse d'une double signification du point de vue épistémologique. Premièrement, elle nous indique que la théorie du traitement de l'information relève des mêmes outils statistiques que ceux qui ont été développés pour la physique. On doit donc s'attendre à une transposition des résultats obtenus en physique dans le domaine du traitement du signal. C'est déjà ce que Wiener a fait lorsqu'il a réutilisé, par exemple, ses travaux sur le mouvement brownien dans ses recherches de guerre sur la théorie de la prédiction et des filtres. Mais à un second niveau, cette inclusion signifie plus qu'une simple relation de dépendance instrumentale entre les concepts physiques et les concepts de la théorie de l'information. Elle implique une interprétation en dernier ressort physicaliste du traitement de l'information. Au niveau fondamental, le traitement de l'information doit être envisagé comme un processus intégralement physique. On trouve ici un des nœuds du matérialisme radical des cybernéticiens. Le recours au concept de traitement de l'information ne sert ni une réinterprétation idéaliste de la physique du point de vue de l'observateur, comme chez Brillouin, ni une étude purement logique des processus de traitement de l'information, comme en intelligence artificielle. Le hardware compte et ne s'efface pas devant le monde idéal du software et des opérations symboliques.

De cette physique de l'information la cybernétique n'a cependant pas retenu, hors quelques résultats marginaux, et, qui plus est erronés, le projet d'une thermodynamique du calcul à la manière de Landauer ou Bennett. C'est sans doute la théorie des automates de Von Neumann qui reflète le mieux l'interprétation originale du concept d'information au sein de la cybernétique. La théorie des automates possède une situation épistémolo-

gique remarquable. D'un côté, il s'agit, en un sens, d'une théorie formelle, et purement formelle, du traitement de l'information. La théorie repose, en effet, largement sur la mise entre parenthèses des conditions physiques effectives, le recours à des composants dont le comportement est idéalisé et qui sont traités comme des « boîtes noires ». Cependant, de l'autre côté, nous avons une théorie qui s'interdit le recours à un traitement de l'information simplement symbolique et qui cherche au contraire à rendre compte de la possibilité de ce traitement symbolique à partir d'un niveau plus fondamental. En ce sens, il y a la volonté manifeste d'expliquer les conditions de possibilité du traitement symbolique de l'information tel qu'il est pratiqué en informatique ou postulé dans les sciences cognitives. Enfin, la théorie des automates, si elle se développe comme une théorie formelle, vise toujours en dernière instance, chez Von Neumann, à rendre compte du traitement effectif, réel de l'information, en particulier au sein des organismes biologiques. L'automate abstrait, l'automate logique, n'est qu'une passerelle vers l'automate biologique.

La théorie des automates peut donc s'interpréter comme la poursuite, par d'autres moyens, du programme physicaliste inscrit dans la relation information-entropie. Nous avons affaire à une élaboration théorique qui propose une forme d'interface entre les différentes dimensions du concept cybernétique d'information. La théorie des automates fait se rejoindre dans une même trame l'information de l'informatique, l'information de la théorie des télécommunications, de la logique ou de la biologie. Dans son dernier texte, inachevé, sur la théorie des automates, Von Neumann écrit que la théorie occupe « une région intermédiaire entre la logique, la théorie de la communication et la physiologie. »[1] La théorie des automates est un des lieux à partir duquel les forces centrifuges du programme cybernétique peuvent être tenues en bride. Elle constitue à soi seule une petite philosophie de l'information et une épistémologie en acte. Elle est un miroir des engagements cybernétiques.

Le développement de la théorie des automates

On peut repérer quelques grands moments dans le développement de la théorie des automates, qui voient apparaître des directions de recherches ou des thèmes nouveaux. Ce que Von Neumann appelle théorie des automates recouvre essentiellement quatre grands champs de questions. Nous avons affaire, premièrement, à une interrogation sur la portée de l'analogie entre le cerveau et l'ordinateur. La théorie des automates fournit un appa-

1. Arthur Burks, *Theory of Self-reproducing Automata*, Urbana, University of Illinois Press, 1966, p. 91.

reil critique pour évaluer la pertinence de l'analogie et surtout tirer des résultats positifs des points sur lesquels celle-ci échoue. Cette question traverse toute l'œuvre. Elle est présente dans l'ensemble des interventions, jusqu'aux dernières conférences Silliman, rassemblées dans *L'Ordinateur et le cerveau*.

Deuxièmement, Von Neumann parvient à démontrer que certains automates possèdent une capacité d'autoreproduction. La question de l'autoreproduction représente une première approche du concept de complexité. À partir d'un certain seuil, on peut s'attendre à ce que certains automates soient en état de produire des automates au moins aussi complexes qu'eux. Cette démonstration de la capacité d'autoreproduction constitue donc le premier grand résultat de la théorie des automates. Il est présenté publiquement pour la première fois lors du symposium Hixon de Pasadena au mois de septembre 1948.

Troisièmement, à partir de 1949, une nouvelle question apparaît, qui prend la forme d'une théorie statistique des automates. Von Neumann cherche à préciser le degré de fiabilité que l'on peut attendre d'un automate composé d'éléments instables. Cette question reçoit un traitement statistique qui constitue une première rupture avec les ressources logiques employées jusqu'ici par la théorie.

Enfin, à partir de la fin de l'année 1952, Von Neumann introduit un nouveau modèle mathématique, celui des automates cellulaires, pour formaliser les questions précédentes. Il donne notamment une preuve d'existence de la construction d'un automate cellulaire autoreproducteur. La théorie des automates se développe donc en intégrant progressivement ces quatre directions de recherche : l'analogie ordinateur-cerveau, l'autoreproduction, le problème de la fiabilité à partir de composants instables, le modèle des automates cellulaires.

Les neurones de McCulloch et Pitts dans le First Draft of a Report on the Edvac

Le premier texte, dans lequel apparaissent les rudiments de ce qui deviendra la théorie des automates, est sans doute le « First Draft of a Report on the Edvac » du 30 juin 1945. Le design logique de l'ordinateur y est défini à partir de composants idéaux, analogues aux neurones de McCulloch et Pitts, avec leurs entrées inhibitrices et excitatrices, leurs sorties et leurs fonctions de seuil. Le « First Draft of a Report on the Edvac » introduit ainsi une première ébauche de la théorie des automates qui repose sur le « vocabulaire » de McCulloch et Pitts. La théorie des automates désignera la théorie de ces composants idéaux qui permettent, au moins dans un premier temps, de masquer la complexité physique inutile

des composants réels. La théorie a cette propriété intéressante qu'elle surplombe le partage entre le vivant et la machine. Vivants comme machines, automates biologiques comme automates logiques, sont susceptibles d'un même type de description, en termes de traitement de l'information.

Mais le « First Draft » reste encore au seuil des décisions épistémologiques qui constitueront la théorie des automates. Il s'agit d'un premier pas, qui ne présente pour l'instant pas autre chose qu'une théorie abstraite du traitement de l'information. L'usage du formalisme de McCulloch et Pitts sert initialement à masquer la complexité physique au profit de considérations purement logiques. Cette première étape partage avec l'option fonctionnaliste le même principe d'indifférence vis-à-vis des réalisations matérielles. La physique de l'information se trouve rejetée à l'arrière-plan comme une entrave inutile, susceptible de compliquer le dégagement du niveau proprement logique du traitement de l'information.

Mais toute la position épistémologique de la théorie des automates va se jouer dans la modification du rôle assigné au formalisme de McCulloch et Pitts. Dès le « First Draft of a Report on the Edvac », nous avons un indice de la façon dont la première grammaire de la théorie des automates pourra être réinterprétée. La théorie abstraite des automates fondée sur le modèle de McCulloch et Pitts est en effet présentée comme une théorie incomplète. Il s'agit, nous dit Von Neumann, d'une théorie provisoire avant qu'on ne discute du design effectif des composants.

> « La procédure idéale consisterait à traiter les éléments comme ce qu'ils doivent être : autrement dit, des tubes à vide. Cependant, cela exigerait une analyse détaillée de questions d'ingénierie radio, à un stade peu avancé de la discussion où trop d'alternatives sont encore ouvertes pour être traitées de manière exhaustive et en détail. Aussi, les nombreuses possibilités alternatives pour construire les procédures arithmétiques, le contrôle logique, etc, se superposeraient aux possibilités également nombreuses quant au choix des types ou des tailles de tubes à vide et autres éléments du circuit du point de vue des performances pratiques, etc. […] Cette simplification est seulement temporaire, elle n'est qu'un point de vue transitoire, afin de rendre cette discussion préliminaire possible. »[1]

La théorie se présente donc comme une simplification en attente d'une théorie physicaliste complète. Non seulement, la théorie n'a pas vocation à se substituer à la théorie complète, mais elle intégrera progressivement une remise en cause des éléments simples de McCulloch et Pitts. Un des champs de recherche de la théorie des automates portera précisément sur les améliorations à apporter au fonctionnement des composants de base pour les rendre plus conformes à ce que l'on sait des composants réels. Le recours au modèle de McCulloch et Pitts, conçu comme modèle incomplet,

1. John Von Neumann, « First draft of a report on the Edvac », *op. cit.*, pp. 8-9.

ouvre ainsi un champ de questions qui porte sur la complexification du comportement des composants de base. Les boîtes noires sont destinées à être ouvertes.

Tout le mouvement de la théorie des automates, dès lors qu'elle se développera pour elle-même, au-delà de la mention du modèle de McCulloch dans le « First Draft », consistera en effet à renverser complètement l'interprétation du rôle des composants formels. Ceux-ci ont été introduits au départ, dans l'article de McCulloch et Pitts, pour représenter de manière logique le fonctionnement du cerveau. Von Neumann les réemploie dans le « First Draft » pour représenter logiquement le fonctionnement des ordinateurs. Mais la théorie des automates prendra le problème de manière inverse. Il ne s'agira plus d'interpréter le cerveau ou les machines à partir de la logique, mais bien plutôt d'interpréter la logique à partir du fonctionnement des machines ou du cerveau, et de rendre compte ainsi à un niveau fondamental des possibilités du traitement logique de l'information. Ce qui est pris dans « First Draft » comme un point de départ, le calcul logique au moyen d'éléments idéalisés, apparaîtra, à l'aune du programme complet de la théorie des automates, comme un résultat ou une conclusion. Il s'agit d'essayer de rendre compte, de manière évidemment formelle et logique, des conditions physiques qui nous permettent de développer un discours formel et une logique.

La lettre du 29 novembre 1946 à Wiener

La théorie des automates prend véritablement corps au cours de l'année 1946, alors que Von Neumann est engagé dans le projet de calculateur de l'IAS. Au printemps 1946, il développe les premiers modèles d'automates autoreproducteurs et donne au mois de juin une série de lectures informelles à Princeton, qui formeront le canevas de son intervention au symposium Hixon en 1948. On peut se faire une idée de l'état de la réflexion de Von Neumann à la fin de l'année 1946 grâce à une lettre extraordinaire, adressée à Wiener le 29 novembre, dans l'intention de préparer leur rencontre début décembre. Cette lettre expose les motivations fondamentales de la constitution de ce qui ne s'appelle pas encore la théorie des automates. La lettre à Wiener constitue un exemple parfait de « rupture épistémologique ». Elle consiste en une discussion sur les programmes de recherches. Ce genre de discussion comporte toujours nécessairement un élément philosophique et implique une discussion qui fait retour sur le sens des concepts engagés. Ici, les concepts d'information et de traitement de l'information. Il s'agit sans doute d'un des moments clés de la première cybernétique.

La lettre nous donne une indication chronologique intéressante. Von Neumann mentionne qu'il a conçu depuis quasiment un an des doutes

quant à la poursuite du programme originel de McCulloch et Pitts. La représentation abstraite du traitement de l'information telle qu'on la trouve dans le « First Draft » de juin 1945 aura finalement duré peu de temps. La théorie des automates apparaît comme le résultat de cette année de doutes qui voit Von Neumann réaménager radicalement l'usage qu'il avait fait du formalisme de McCulloch et Pitts. Von Neumann explique avoir tenu ses doutes en réserve faute d'un véritable programme alternatif. La théorie des automates s'impose comme cette alternative au sein du premier programme cybernétique.

En quoi consiste précisément le message de la lettre ? Von Neumann presse Wiener de changer de tactique. Jusqu'ici les cybernéticiens se sont intéressés surtout au fonctionnement du système nerveux central, en particulier humain, autrement dit à un objet de très haute complexité. En dépit de cette complexité, ils ont pu obtenir un résultat remarquable. Turing nous a appris qu'il existe des mécanismes universels capables de calculer n'importe quelle fonction calculable de premier ordre, et McCulloch et Pitts nous ont appris que le cerveau humain possédait les propriétés minimales d'un tel calculateur universel. Mais une fois ces résultats d'une très grande généralité assimilés, que peut-on montrer de plus, si ce n'est au moyen d'une véritable étude microscopique du système nerveux ?

> « Ce sur quoi il me semble important d'insister est qu'une fois la grande contribution de Turing-cum-Pitts-et-McCulloch assimilée, la situation est plutôt pire qu'avant. En effet, ces auteurs ont démontré avec une généralité absolue et désespérante que toute la logique de Brouwer peut être effectuée par un mécanisme approprié, et en particulier un mécanisme neuronal – et que même un mécanisme particulier peut être " universel ". Inversons l'argument : rien de ce que nous pouvons savoir ou apprendre au sujet du fonctionnement de l'organisme ne peut nous donner, sans un travail " microscopique ", cytologique, aucune clé quant aux détails avancés du mécanisme neuronal. »[1]

Ainsi, les seuls progrès véritables à faire dans l'étude du système nerveux ne se situent plus du côté de la logique – une fois le résultat de McCulloch et Pitts acquis –, mais du côté de l'étude neurologique fine. Or le système nerveux est un objet d'une incroyable complexité qui échappe encore largement à nos méthodes d'investigation biologiques. Non seulement le nombre de composants est énorme, mais nombre de processus ont lieu de manière analogique et non digitale. Autrement dit, l'étude de détail du système nerveux est beaucoup trop complexe au regard des moyens actuels. Essayer de comprendre le système nerveux avec les moyens actuels serait

1. John Von Neumann à Norbert Wiener, 29 Novembre 1946, in Miklos Rédei (dir.), *John Von Neumann, Selected Letters*, American Mathematical Society, 2005, p. 278.

comme « essayer de comprendre l'ENIAC sans jamais avoir entendu parler d'arithmétique. »[1] Étudier des systèmes nerveux plus simples que le système nerveux humain, comme celui des fourmis par exemple, ne change rien à l'affaire, argumente Von Neumann, car la complexité que l'on gagne en termes de composants se perd de l'autre côté : la part analogique augmente et devient moins accessible, on ne peut plus s'appuyer sur la connaissance des pathologies et nos possibilités de communication sont diminuées…

Von Neumann suggère alors un changement de cap radical consistant à s'attaquer non plus à des organismes complexes, mais aux composants de base des systèmes biologiques. Abandonner par exemple une étude du cerveau, qui postule le fonctionnement stable et logique du neurone, pour s'intéresser au neurone lui-même et rendre compte de ses performances réelles. Autrement dit, c'est toute la tâche assignée au formalisme de la théorie des automates qui se trouve renversée. En 1945, dans le « First Draft of Report on the Edvac », le formalisme de McCulloch et Pitts servait à masquer le fonctionnement des composants en faveur d'une interprétation purement logique et fonctionnaliste de la machine. En 1946, le formalisme de McCulloch et Pitts doit servir, non plus à masquer le fonctionnement des composants, mais à l'étudier et à le décrire en tant que tel. C'est toute l'axiomatique qui se trouve la tête à l'envers.

Von Neumann emploie une analogie technologique pour exposer la nécessité de cette authentique révolution, d'où procède la théorie des automates.

> « Cela doit nous rendre suspect le choix des cellules comme concepts de base " non définis " d'une axiomatique. Pour être plus terre-à-terre : considérons, dans n'importe quel champ de la technologie, le moment qui se caractérise par le développement de " composants standards " hautement complexes, qui sont dans le même temps individualisés, bien adaptés à la production de masse, et (en dépit de leur caractère " standard ") bien adaptés à différents usages. Ceci représente clairement un style tardif et hautement développé, et non celui qui est idéal pour une première approche du sujet de la part de quelqu'un d'extérieur. »[2]

Qu'est-ce que ce renversement vis-à-vis du statut des composants implique au point de vue épistémologique ? L'usage des neurones formels de McCulloch et Pitts qui est à la base de la théorie des automates change profondément de signification à partir de l'année 1946. Ils ne servent plus à asseoir une interprétation logique du fonctionnement du cerveau, mais ils servent à décrire la logique des comportements de base du vivant. Il ne s'agit donc plus de montrer que le cerveau est formellement capable d'une performance cognitive comme le calcul des propositions, mais de s'attaquer

1. John Von Neumann à Norbert Wiener, 29 Novembre 1946, in Miklos Rédei (dir.), *John Von Neumann, Selected Letters*, *op. cit.*, p. 279.
2. *Ibid.*, pp. 279-280.

à la compréhension des performances biologiques infra-cognitives des cellules, des virus, des bactériophages ou des enzymes, pour reprendre les exemples donnés par Von Neumann. La théorie des automates se présente ainsi comme une nouvelle logique du vivant. Mais ce faisant, c'est tout le statut de la logique qui a changé de sens pour se lier de manière intime à la description du comportement des organismes. Il ne s'agira plus de démontrer que tel ou tel système est capable d'une performance logique, au sens ordinaire, comme le cerveau peut implémenter la logique du premier ordre dans la théorie de McCulloch et Pitts, mais de caractériser d'une manière logique et formelle le comportement d'un système. Nous avons affaire à une logique qui est toujours investie dans le comportement d'un système matériel. Le traitement de l'information abandonne le royaume des programmes ou de la logique abstraite pour descendre s'immiscer dans les interstices du comportement des vivants les plus simples. On pourra alors s'interroger sur le nombre de composants simples requis pour telle ou telle performance, sur le type d'organisation formelle nécessaire, etc. Von Neumann suggère ainsi plusieurs champs de recherche qui concernent des performances fondamentales comme l'autoreproduction, l'orientation dans un milieu, le maintien du milieu intérieur… Et il annonce avoir déjà obtenu ses premiers résultats sur le terrain de l'autoreproduction.

> « Les organismes moins-que-cellulaires du type du virus ou du bactériophage possèdent les traits décisifs des organismes vivants : ils sont autoreproducteurs et ils sont capables de s'orienter d'eux-mêmes dans un milieu inorganisé, de se déplacer en direction de la nourriture, de se l'approprier et de l'utiliser. En conséquence une “ véritable ” compréhension de ces organismes peut constituer le premier pas en avant et peut-être le plus grand pas qu'on puisse exiger. […] J'ai pas mal réfléchi au sujet des mécanismes autoreproducteurs. Je peux formuler le problème de manière rigoureuse, à peu près dans le même style que Turing pour ses mécanismes. Je peux montrer qu'ils existent dans ce système de concepts. Je pense que je comprends certains des principes fondamentaux qui sont impliqués. Je veux compléter les détails et rédiger ces considérations au cours des deux prochains mois. J'espère apprendre diverses choses au cours de cet exercice littéraire, en particulier, le nombre de composants requis pour l'autoreproduction. »[1]

La lettre se conclut sur l'évocation enthousiaste de la « prochaine rupture décisive » qui doit résulter de cette réorientation complète de la théorie des automates. À l'analyse des procédés d'autoreproduction, Von Neumann ajoute déjà l'étude de la relation gène-enzyme ou des mutations. Le canevas général de la théorie des automates est ainsi établi dès la fin de l'année 1946, en pleine période d'effervescence cybernétique, au moment où Von Neumann réoriente le modèle de McCulloch et Pitts vers l'étude des propriétés fondamentales de l'organisation biologique.

1. John Von Neumann à Norbert Wiener, 29 novembre 1946, in Miklos Rédei (dir.), *John Von Neumann, Selected Letters*, *op. cit.*, pp. 280-282.